УДК 004.891.3

ОТБОР ИНФОРМАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧАХ АНАЛИЗА И КЛАССИФИКАЦИИ СУДЕБНЫХ ДОКУМЕНТОВ

А.А. Алексеев, А.С. Катасёв,
Казанский национальный исследовательский
технический университет им. А.Н. Туполева-КАИ;
А.Е. Кириллов,
Арбитражный суд Республики Татарстан, Россия, г. Казань

Ключевые слова: *классификация текстовых документов, судебный акт, терм-документная матрица, статистическая мера TF-IDF, энтропия.*

В настоящее время Арбитражным судом Республики Татарстан в год принимается 25–30 тысяч судебных решений, которые относятся к 38 категориям споров, таким как преддоговорные споры, взыскание задолженности, оспаривание решений административных органов и др. С целью повышения эффективности судебного документооборота необходима разработка механизма автоматической классификации судебных актов в соответствии с категориями споров. Данный механизм позволит в автоматическом режиме определять категорию и сложность судебных актов для их более равномерного распределения между отделами, а также улучшить сбор и обработку статистических данных о работе суда и исключить в статистической отчетности недостоверных сведений о категориях рассмотренных дел.

Согласно [1], большинство методов классификации текстов, так или иначе, основаны на предположении, что документы, относящиеся к одной категории, содержат одинаковые признаки (слова или словосочетания). Наличие или отсутствие таких признаков в документе говорит о его принадлежности или непринадлежности к той или иной категории. В данной работе для представления текста в форме, удобной для анализа, используется терм-документная матрица (term-document matrix – TDM), представляющая собой таблицу, где каждая строка соответствует термину, а столбец – документу в наборе данных [2]. На пересечении строк и столбцов хранятся значения весов терминов в документе. Каждый вес представляется статистической мерой *TF-IDF* [3], позволяющий выделить наиболее важные для определенного набора документов термины:

$$TF * IDF = \frac{n_t}{\sum_k n_k} * \log \frac{D}{(t \in d_i)}, \qquad (1)$$

где n_t – количество раз встречаемого слова t документа d, $\sum_k n_k$ – длина данного документа (количество слов в документе), D – общее количество документов, $t \in d_i$ – количество документов, в которых содержится слово t.

Термины формируются на этапе предварительной обработки, где основными приемами являются токенизация текста, фильтрация стоп-слов, стэмминг (лемматизация), приведение регистра.

В качестве примера рассмотрим 4 категории судебных споров: оспаривание действий судебных приставов, оспаривание решений антимонопольных органов, привлечение к ответственности за нарушение условий поставок, привлечение к ответственности за нарушение условий лицензирования. В исходном множестве документов первая категория содержит 37 судебных акта, вторая – 32, третья – 63, четвертая – 36. На этапе предварительного анализа сформирована терм-документная матрица размерностью 3996*168, т. е. 3996 полученных терминов в 168 документах. Фрагмент матрицы представлен в табл. 1.

Таблица 1

Фрагмент терм-документной матрицы

Класс / Термин	Class_1	Class_2	Class_3	Class_4
Антиконкурентн	0	0.0037	0	0
Антикризисн	0.0095	0	0	0
Антимонопольн	0.0057	0.5934	0	0
Апелляцион	0.0059	0.0025	0.0102	0.0057

Среди 3996 терминов необходимо выявить наиболее значимые для своей категории, так как подача всего их числа в качестве входных данных на классификаторы вызовет затруднения при вычислении в виду временных затрат на обучения данных классификаторов и загрузки памяти. Для выделения наиболее информативных признаков в данной работе применяется два подхода. Первым из них является статистическая мера *TF-IDF*. В данном подходе оба множителя, *TF* и *IDF*, имеют значение при подсчете веса термина во всем множестве документов. Так как *TF* является частотой термина в документе, то чем чаще термин встречается в одном акте, тем значение *TF* выше. Таким образом, *TF* играет роль повышающего множителя в анализе. Однако возникает необходимость отсеять термины, которые встречаются в каждой категории судебных актов и, соответственно, имеют большое значение *TF*. Для отсеивания применяется второй множитель *IDF*, являющийся понижающим. Чем больше документов, содержащих один и тот же термин, тем значение *IDF* ниже.

При построении терм-документной матрицы, столбцы, соответствующие документам, были объединены по своим категориям, а значения в них просуммированы по каждому термину. Такой подход наглядно позволяет выявить наиболее информативные термины для каждого из 4 классов (см. табл. 2).

Таблица 2

Значения *TF-IDF* терминов, характеризующих категории судебных документов

Термин	*TF-IDF* класса 1	Термин	*TF-IDF* класса 2	Термин	*TF-IDF* класса 3	Термин	*TF-IDF* класса 4
пристав	1.436	заказ	0.642	договор	1.174	пассажир	0.967
пристава исполнител	0.465	антимонопольн	0.593	истц	0.947	автобус	0.727
взыскател	0.4	уфас	0.211	накладн	0.622	перевозок	0.724
арест	0.386	проект	0.21	неустойк	0.513	маршрут	0.531

Вторым подходом при отборе информативных признаков выбрано вычисление взаимной информации через энтропию [4]:

$$H(X) = -\sum_{x_i \in X} p(x_i) * \log_2(p(x_i)), \quad (2)$$

где $p(x_i)$ – вероятность того, что переменная X примет значение x_i. В сформированной терм-документной матрице данная вероятность определяется как отношение числа вхождений термина в документы к общему количеству терминов.

Взаимная информация определяется следующим образом:

$$I(Y|X) = H(Y) - H(Y|X), \quad (3)$$

где $H(Y|X)$ – условная энтропия:

$$H(Y \mid X) = -\sum_{x_i \in X} p(x_i) * H(Y \mid X = x_i), \quad (4)$$

где $H(Y|X=x_i)$ – частная условная энтропия $H(Y)$ относительно отдельного термина x_i. При отборе признаков наиболее значимыми являются те, что имеют наибольшее значение взаимной информации $I(Y|X)$ [5].

В табл. 3 представлены значимые термины, определенные при помощи взаимной информации.

Таблица 3

Значения взаимной информации терминов, характеризующих категории судебных документов

Термин	Взаимная информация класса 1	Термин	Взаимная информация класса 2	Термин	Взаимная информация класса 3	Термин	Взаимная информация класса 4
пристав	0.029	заказ	0.014	истц	0.035	деятельн	0.013
приста-ваисполни-тел	0.01	антимоно-польн	0.02	рубл	0.033	пассажир	0.013
должник	0.019	размещен	0.016	обяза-тельств	0.022	перевозок	0.01
исполни-тельн	0.049	конкуренц	0.015	накладн	0.02	транспорт	0.01

Представленные выше методы актуальны при решении задач классификации, кластеризации и поиска подобных текстовых документов. Термины, отобранные при помощи данных методов, позволяют эффективно определить категорию судебного документа. При подаче на вход классификатора, построенного на базе нейронной [6, 7] или нечеткой нейронной сети [8, 9], будет задействовано всего 16–20 параметров, а не 3996, как в изначальной терм-документной матрице.

Литература

1. Барсегян А.А. Анализ данных и процессов: учеб. пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод, М.Д. Тесс, С.И. Елизаров. – 3-е изд., перераб. и доп. – СПб.: БХВ-Петербург, 2009. – 512 с.: ил.

2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press New York, NY, USA © 2008. – 496 p.

3. Ingo Feinerer, Kurt Hornik, David Meyer. Text Mining Infrastructure in R. Journal of Statistical Software. March 2008, Volume 25, Issue 5. – 54 p.

4. Ризаев И.С., Ляшева С.А., Шлеймович М.П. Теория информации: учебное пособие. Казань: Изд-во Казан. гос. техн. ун-та, 2008. 88 с.

5. Трегубов В.М., Катасёв А.С., Кириллов А.Е., Алексеев А.А. Информационная технология анализа и классификации электронных документов / Поиск эффективных решений в процессе создания и реализации научных разработок в российской авиационной и ракетно-космической промышленности. Международная научно-практическая конференция. Казань. – 2014. – С. 345–348.

6. Корнилов Г.С., Аникин И.В., Катасёв А.С. Методы и алгоритмы преднастройки и оптимизации параметров нечеткой нейронной сети // Международная конференция по мягким вычислениям и измерениям. – 2009. – Т. 1. – С. 223–226.

7. Катасёв А.С., Катасёва Д.В., Кирпичников А.П. Нейросетевая технология классификации электронных почтовых сообщений // Вестник Казанского технологического университета. – 2015. – Т. 18. – № 5. – С. 180–183.

8. Катасёв А.С. Математическое и программное обеспечение формирования баз знаний мягких экспертных систем диагностики состояния сложных объектов / Монография. – Казань, 2013. – 200 с.

9. Катасёв А.С. Нейронечёткая модель и программный комплекс формирования баз знаний экспертных систем // Диссертация на соискание ученой степени кандидата технических наук. – Казань, 2006.